DISCRIMINABILITY DISTILLATION IN GROUP REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning group representation is a commonly concerned issue in tasks where the basic unit is a group, set or sequence. The computer vision community tries to tackle it by aggregating the elements in a group based on an indicator either defined by human such as the quality or saliency of an element, or generated by a black box such as the attention score or output of a RNN.

This article provides a more essential and explicable view. We claim the most significant indicator to show whether the group representation can be benefited from an element is not the quality, or an inexplicable score, but the *discrimiability*. Our key insight is to explicitly design the *discrimiability* using embedded class centroids on a proxy set, and show the discrimiability distribution *w.r.t.* the element space can be distilled by a light-weight auxiliary distillation network. This processing is called *discriminability distillation learning* (DDL). We show the proposed DDL can be flexibly plugged into many group based recognition tasks without influencing the training procedure of the original tasks. Comprehensive experiments on set-to-set face recognition and action recognition valid the advantage of DDL on both accuracy and efficiency, and it pushes forward the state-of-the-art results on these tasks by an impressive margin.

1 INTRODUCTION

With the rapid development of deep learning and the easy access to large-scale group data, recognition tasks using group information has drawn great attention in the computer vision community. The rich information provided by different frames can complement each other to boost the performance of tasks such as face recognition, action recognition and person re-identification. While common practice is to either aggregate the whole set by average or max-pooling, or sample randomly from a whole video, they ignore the fact that certain frames contributes negatively towards recognition tasks. Thus, an important issue is to select represent samples from the whole set efficiently for group understanding.

To tackle such cases, previous methods aims at defining quality or saliency of an element, or learning weights automatically in a self-attentional manner. For example, Liu et al. (2017b) propose the Quality Aware Network (QAN) to learn quality scores for each frame inside an image set during network training. Other works borrow same idea and extend to specific tasks such as video-based person re-identification by learning spatial-temporal attentions. However, the whole online quality or attention learning procedure are either manually designed or learned through a black box, which lack explainablility. Also, their online learning has to be specifically trained with the main network, which cost great computation burden.

In this work, we explore deeper into the underlying mechanism for defining effective elements instead of relying on self-learned attention. Assuming that a base network has already been trained for element-based recognition using class labels, we define the "discriminability" of one sample by how difficult it is for the network to discriminate its class. It can be observed that the feature embedding of elements lie close to the centroid of its corresponding class are the representatives of this class, while features far away or closer to other classes are the confusing ones which are not distinguishable enough. Inspired by this observation, we identify a successful discriminability indicator by *measuring one embedding's cosine distance with class centroids, and compute the ratio of between positive and hardest-negative,* where the positive is its distance with its class's corresponding centriod and the hardest-negative is the closest counterpart. This process is defined as discriminability distillation regulation (DDR).

Armed with recent theories on the homogeneity between class centroids and projection weights of classifiers, the entire distance-measuring procedure can be easily accomplished by simply encoding all elements in one group. Thus, discriminability scores can be assessed for each element after the training of the base network. This assessing procedure is highly flexible without quality supervision nor re-training base network, so it can be adopted to any existing base. With our explicitly designed discriminability indicator on the training set, the distillation of such discriminability can be successfully performed with a light-weight network, which shows the superiority of our proposed indicator. We call our whole procedure uniformly as *discriminability distillation learning* (DDL).

The next step is towards finding better aggregation policy. At the test phase, all samples are first sent to the light-weight discriminability distillation network, only frames will high scores will be weighted and aggregated. We evaluate the effectiveness of our proposed DDL on two classical yet challenging tasks: set-to-set face recognition and action recognition. Comprehensive experiments show the advantage of our method on both recognition accuracy and computational efficiency. We achieve state-of-the-art results without modifying the base networks.

We highlight our contributions as follows: (1) We define the *discriminability* of one element within a group from a more essential and explicable view. (2) We verify that a light-weight network has the capacity of distilling discriminability from the assessed elements. Combining the post-processing with the network, great computational burden can be saved comparing with existing methods. (3) We validate the effectiveness of DDL for both efficiency and accuracy on set-to-set face recognition and action recognition through extensive studies. State-of-the-art results can be achieved.

2 RELEATED WORK

2.1 SET TO SET FACE RECOGNITION

Convolutional neural network has achieved great success in face recognition. Many new loss functions Deng et al. (2019a); Schroff et al. (2015); Liu et al. (2017a); Sun et al. (2014); Wen et al. (2016), training datasets Cao et al. (2018); Yi et al. (2014); Wang et al. (2018); Guo et al. (2016), and neural archte He et al. (2016); Zhang et al. (2017); Chen et al. (2018) are proposed to boost face recognition performance. The accuracy achieved by state-of-the-arts face recognition model ArcFace Deng et al. (2019a) on a million scale still image test benchmark MegaFace has been 98.35%. However, in real-world face verification application, it is more practical to recognize a face with adjacent video frames in a set to set manner.

The core problem of the set to set recognition task is face feature aggregation. Since not all frames have salient information and some frame meet with visual blur and large pose problems, simply average or max-pooling feature performs not well. There are works proposed to optimize set to set recognition with attention or quality mechanism. Yang et al. (2017) uses two cascaded attention module to aggregate frame feature. While QAN Liu et al. (2017b) proposes a quality-aware side-network trained with a specifically designed main network. During testing, the video-level feature is weight aggregated by quality. Our work is inspired by QAN but introduces a robust post-training mechanism which decouple quality network and base recognition model. State-of-the-art results can be achieved.

2.2 ACTION RECOGNITION

With the advence of the multimedia era, millions of hours of video are uploaded to video platforms every day, so video understanding task like action recognition has become a popular research topic. There are two typical approaches to action recognition. The first one is to extract feature by frames and use sophisticated late fusion strategy to form video-level prediction Yue-Hei Ng et al. (2015); Simonyan & Zisserman (2014a); Girdhar et al. (2017). However, aggregate high-level frame feature tends to limit the model ability to handle complex motion and temporal relation.

Other approaches are to use 3D convolutional neural network to jointly capture spatio-temporal features, which perform well in action recognition. Tran et al. (2015) first proposes 3D convolu-

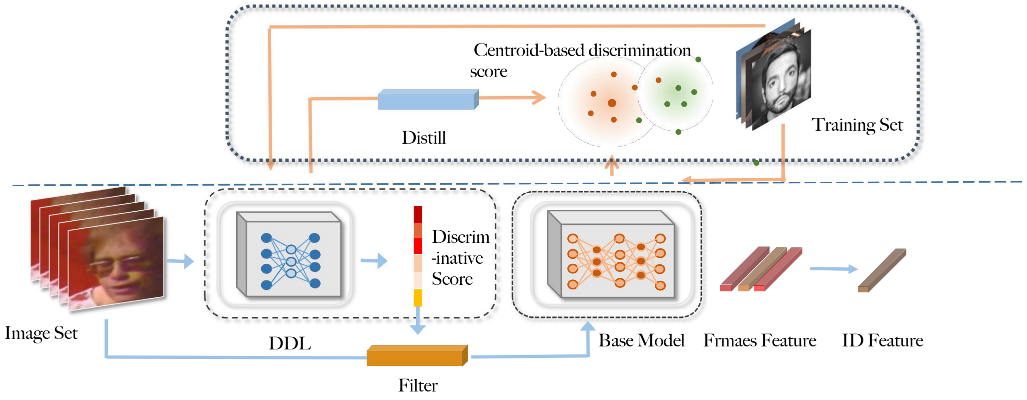

Figure 1: The pipeline of group representation learning with our DDL. Given a base feature extract model, we first compute the discriminability score for each training image and then trianing a lightweight CNN $\mathcal{N}$ to approximate it. For testing, all frames are first sent to $\mathcal{N}$ to approximate the discriminability score and then the group representation can be generated by well-designed feature aggregation module.

tion neural network for video recognition based on an VGG Simonyan & Zisserman (2014b) like network. Tran et al. (2018) proposes R(2+1)D convolution, which separates 3d convolution to 2d convolution following by a 1d convolution. R(2+1)D introduces more non-linear operation and reduce the risk of over-fitting. SlowFast Feichtenhofer et al. (2018) proposes a two patch 3D convolution architecture and achieves state-of-the-art performance. Though the above methods achieve huge success in action recognition, they need a large number of computing resources due to dense sample strategy during inference, making it unpractical to application.

2.3 Distillation

Our approach is also inspired by distilling teacher networks into student models Hinton et al. (2015). However, traditional distillation usually trains the student model with the same task with the teacher model for the purpose that achieves better recognition performance for the flops-constrained model. Our DDL introduces a quality distillation mechanism, the light-weight network are designed to distill the centroid-based quality measured by the base model prediction confidence and help the base model to achieve accurate and efficient recognition in a frame set manner.

3 Discriminability Distillation Learning

In this paper, we aim to explore the discriminative score of an image in a given image set. Beyond the traditional quality learning strategies where the quality score of image is learned by well-designed attention mechanism or manual annotation, we propose to distill the discriminative score from the feature space distribution. Based on the interpretable and reasonable distillation rules, the discriminative score is practical for learning group representation and it can be easily plugged into the popular group representation learning framework. In the following subsections, we first formulate the group representation and then define the discriminability distillation regulation. Finally, the whole discriminability distillation learning (DDL) is introduced in subsection 3.3.

3.1 Formulation of Group Representation Learning

Group representation learning focuses on formulating an uniform representation for a whole set of images. Whether verification task or classification task, the core of them is how to aggregate the features of a given image set.

Define the f_i as the embedded feature of frame I_i in an image set I_S, the aggregated feature F_{I_S} is generated by:

$$F_{I_S} = \mathcal{G}(f_1, f_2, \cdots, f_i) \tag{1}$$

Where $\mathcal{G}$ indicates the feature aggregation module. In the previous research, conducting $\mathcal{G}$ with quality score has become a priority. Different from the general quality learning via attention mechanism or manual annotation, we propose a discriminability distillation learning (DDL) to genearte the *discriminability* of a feature representation. Furthermore, the well-designed discriminability distillation regulation in DDL makes it interpretable and reasonable.

3.2 Discriminability Distillation Regulation

Towards efficient and accurate $\mathcal{G}$, we design the discriminability distillation regulation (DDR) to generate the *discriminability* to replace the traditional quality score. In DDR, we joinly consider the feature space distribution and explicitly distill the *discriminability* via encoding the intra-class distance and inter-class distance with class centroids. Define the training dataset $\mathcal{X}$ with C identities and the class centroids $W_j, j \in [1, C]$ in the final classification layer, for feature $f_i, i \in [1, s]$ with identity c where s indicates the length of $\mathcal{X}$, the intra-class distance and inter-class distance are formulated as:

$$\begin{aligned} \mathcal{C}_{ia} &= \frac{f_i \cdot W_c}{\|f_i\|_2 \; \|W_c\|_2} \\ \mathcal{C}_{ie} &= \frac{f_i \cdot W_j}{\|f_i\|_2 \; \|W_j\|_2}, \; j \neq c \end{aligned} \tag{2}$$

By jointly considering this, we define the *discriminability* Q_i of f_i via the DDR as:

$$Q_i = \frac{\mathcal{C}_{ia}}{\max\{\mathcal{C}_{ie} \mid j \in [1, C], j \neq c\}} \tag{3}$$

Considering the variant number of images in different groups, we further normalize the *discriminability* by:

$$\mathcal{D}_i = \tau\left(\frac{Q_i - \mu(\{Q_j \mid j \in [1, s]\})}{\sigma(\{Q_j \mid j \in [1, s]\})}\right) \tag{4}$$

where $\tau(\cdot)$, $\mu(\cdot)$ and $\sigma(\cdot)$ mean the sigmoid function, mean value and standard deviation value of $\{Q_j \mid j \in [1, s]\}$, respectively.

Coorperated with the feature space distribution, the *discriminability* $\mathcal{D}_i$ is more interpretable and reasonable than the quality score in traditional quality learning. It can better represent the discriminability of a feature based on explicitly encoding the intra-class distance and inter-class distance with class centroids.

3.3 Discriminability Distillation Learning

According to the Sec 3.2, given the training dataset θ and the corresponding model, the *discriminability* $\mathcal{D}_i$ of f_i can be naturally computed via Eq(2)(3)(4). However, for the test set $\mathcal{T}$ in group representation learning, it's unavailable to the $\mathcal{D}_i$ due to the lack of W_j. In order to better embed the $\mathcal{D}_i$ into the group representation learning, we introduce the discriminability distillation learning (DDL) to generate the approximated *discriminability* $\hat{\mathcal{D}}_i$ of $\mathcal{D}_i$. Given an arbitrary CNN architecture $\mathcal{N}$ and the image I_i, the *discriminability* $\hat{\mathcal{D}}_i$ can be approximated by:

$$\hat{\mathcal{D}}_i = \mathcal{N}(I_i; \boldsymbol{\theta}) \tag{5}$$

where $\boldsymbol{\theta}$ is the parameter of $\mathcal{N}$. At the training stage, we apply mean squared error between each $\hat{\mathcal{D}}_i$ and target $\mathcal{D}_i$ as:

$$L = \frac{1}{2N} \sum_{i}^{N} (\hat{\mathcal{D}}_i - \mathcal{D}_i)^2 \tag{6}$$

where N is the batch size. At the inference stage, $\mathcal{N}$ can assess each frame a *discriminability* $\hat{\mathcal{D}}_i$.

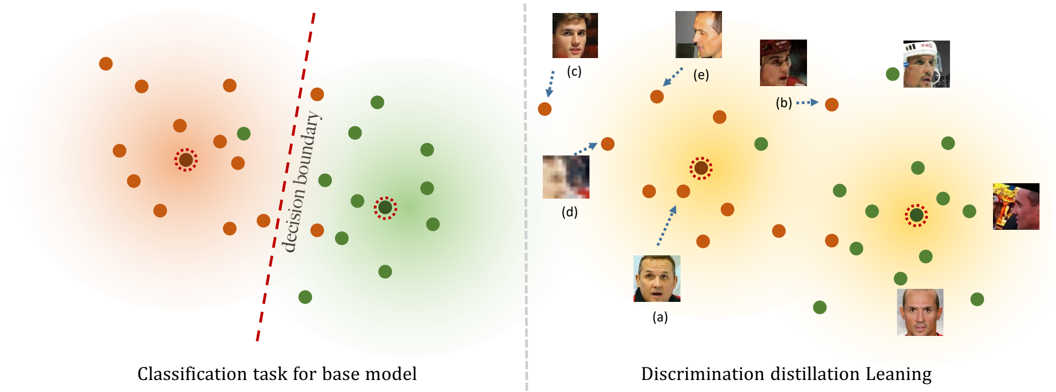
decision boundary
(c)
(e)
(b)
(d)
(a)
Classification task for base model
Discrimination distillation Leaning

Figure 2: The formulation of Discriminability. The base model is trained with classification task, and features of samples from the same class are projected to hyperspace tightly in order to form a clear decision boundary. The outliers are usually not discriminative samples. We list several situations here. (b)appearance similar to neighboring class,(c) data noise (d) visual blur (e) large pose, which for (a), a clear front face, which is discriminative.

3.4 Feature Aggregation $\mathcal{G}$

At the test stage, we can generate the *discriminability* $\hat{\mathcal{D}}_i$ via Eq(5) for each test image I_i in the given image set I_S. The feature aggregation $\mathcal{G}$ in Eq (1) can be formulated as:

$$F_{I_S} = \mathcal{G}(f_1, f_2, \cdots, f_n) = \sum_{i}^{n} \frac{\hat{\mathcal{R}}_i f_i}{\hat{\mathcal{R}}_i} \tag{7}$$

where n is the image number of I_S. $\hat{\mathcal{R}}_i$ is the re-scaled *discriminability* of $\hat{\mathcal{D}}_i$ via:

$$\hat{\mathcal{R}}_i = K\hat{\mathcal{D}}_i + B \tag{8}$$

In Eq(8), we linearly map the minimum *discriminability* and maximum *discriminability* of image set I_S to 0 and 1, respectively. This mapping ensures that image sets with different lengths have the same range of *discriminability*. The K and B are formulated as:

$$K = \frac{1}{\max\{\hat{\mathcal{D}}_i \mid i \in [1, n]\} - \min\{\hat{\mathcal{D}}_i \mid i \in [1, n]\}} \tag{9}$$

$$B = 1 - K \max\{\hat{\mathcal{D}}_i \mid i \in [1, n]\} \tag{10}$$

3.5 Advantage of Discriminability Distillation Learning

Different from the subjective quality judgment of an image or quality learning via attention mechanism, we explicitly assign *discriminability* for each image via the feature space distribution. By jointly considering the inter-class distance and intra-class distrance with class centroids, DDL can effectively approximate how discriminative is a feature. According to aggregate more information from the features with high *discriminability* in an image set, the recognition performance can be easily boosted. With the assistant of DDL, both hard samples (the model easily failed on them) or low-quality image (the model is easily to give wrong predict result) will be filtered. For face recognition, DDL can easily improve the performance by concentrating the discriminative information and for video action recognition, DDL can further accelerate the pipeline by eliminating the frames with insufficient information.

Method	Data	Accuracy(%)
NAN	3M	95.5
Center Loss	0.7M	94.9
FaceNet	200M	95.1
QAN	5M	96.2
UniformFace	6.1M	97.7
CosFace	5M	97.6
SphereFace	0.5M	95.0
ArcFace+Avg	5.1M	98.3
ArcFace+DDL		**99.2**

Table 1: Comparison with different methods on the Youtube Face benchmark. DDL will base model trained with ArcFace loss achieves the state-of-the-art verification performance.

4 Experiments

In this section, we first evaluate our DDL for set to set face recognition on YTF dataset. Next, for action recognition, we conduct experiments on untrimmed video dataset ActivityNet-1.2 and trimmed video dataset Kinetics-700.

Datasets. We train our base recognition model and distillation network on MS-Celeb-1M dataset Guo et al. (2016). Since the original dataset is proved to be dirty, we use the cleaned version by Deng et al. (2019a), which consists of 5,179,510 pictures from 93,431 identities.

Implementation details. RetinaFace Deng et al. (2019b) is used to detect and align training images to 112×112. For data augmentation, only random flip is used. For the base recognition model, the backbone network we select is the modified version of ResNet-101 He et al. (2016) introduced by Deng et al. (2019a). After the last convolution layer, the network outputs a 256-D embedding feature. ArcFace Deng et al. (2019a) function loss is used to form more discriminate feature. The hyperparameter scale and margin are 64 and 0.5 respectively. Learning rate is initially with 0.1 and divides by 10 at 60k, 80k iterations, and we finish our training at 11k. 16 GPUS with total batch size 1024 and Pytorch Paszke et al. (2017) framework are used.

Details of DDL architecture

All frames will be sent to the quality network at the inference stage, so the compute burden for DDL is sensitive. To make the inference stage efficient, we design the distillation network to be light-weight and compute efficiently. DDL is a channel reduce ResNet-18 convolution. Our DDL only introduces 81.9 Mflops computation amount (112*112) input, compared to the base feature extract model which often reaches 24Gflops (ResNet101), it is super-efficient.

4.1 Set to Set face recognition

For DDL, we use the light-weight Resnet-like architecture which has been shown above. Discrimination scores are generated for each image in the training set following the pipeline in section 3. Then DDL will regress them with the same RGB input. L2 loss is used and the training procedure is similar to the base recognition model. During testing, the DDL will predict scores for set images and the set level feature will be aggregated by discrimination score weight average.

Evaluation on YouTube Face.

The YTF dataset includes 3425 videos of 1595 identities with an average of 2.15 videos per identity. The videos vary from 48 frames to 6,070 frames. They are challenging because most faces are blurred and have large variations in pose, expression. The performance test on YTF datasets is evaluated under the unrestricted with labeled outside data protocol. We strictly follow the official test protocol that evaluate using 10-fold cross-validation.

As shown in table **??**, our DDL achieves state-of-the-art performance on the Youtube Face benchmark. What's more, from the comparison for different aggregation strategies like average pooling,

Aggregation	Base model	TPR@FPR=1e-4(%)
Avg	R100	65.843
Top1 Quality		65.217
DDL w/o Re-scale		67.381
DDL		69.048
DDL	PolyNet	**72.981** (rank **1st** in the leaderborad)

Table 2: Comparison with different aggreation stragegy on the IQIYI-VID-FACE challenge. DDL achieve huge performance gain compared to simple avgerage pooling. By combing with strong base model PolyNet, DDL achieves 1st performance in the leaderboard. (http://www.insightface-challenge.com/results)

SF	SF(DDL)	CF	CF(DDL)	AF	AF(DDL)
95.8	97.1	97.8	98.7	98.3	99.2

Table 3: Combine DDL with more baselines including SphereFace (SF) and CosFace (CF). All baseline models are trained with same backbone and datasets. Average pooling is used as the default aggregation strategy.

DDL can significantly boost performance, which indicates DLL has has learned a meaningful pattern for discriminability.

As a post-training module, DDL can cooperate with any pre-trained model, so we re-implement more baseline and combine them with DDL, the results are shown in the table 3. In cooperation with DDL, all baseline models achieve huge performance gain compared to the simple average pooling fusion strategy. Those experiments show our DDL is robust and has good generalization ability.

4.2 Results on iQIYI-VID-FACE

The iQIYI-VID-FACE dataset iQIYI (2019) aims to identify the person in entertainment video by face images. The total dataset contains 643,816 video clips of 10,034 identities. To simplify, face frames are extracted from each video at 8FPS and preprocessed to size 112*112. 6.3 M preprocessed face crops are provided finally.

The test protocol is 1:1 verification and the true positive rate under false positive rate at $1e^{-}4$ is reported. The results are shown in table 3. QAN improve 3.2% TPR compared to average pooling and 1.7% to weight sum, which indicate the superb performance of DDL. By combing DDL with stronger base recognition model trained with PolyNet, our DDL achieves 72.981, which is rank 1st method in the leadboard.

4.3 Video action recognition

Evaluation on ActivityNet-1.2

The discriminability distillation learning is more practical to untrimmed video action recognition since there are more diversity videos chip with ambiguous content and visual blur problem. Moreover, dense sampling clips along with the whole video and do average pooling to aggregate is impossible due to massive computational requirements for untrimmed video. For example, the computational consumption of one 2 minutes long video can reach 95,400 Gflops (224*224 input) for the state-of-the-art action recognition model SlowFast Feichtenhofer et al. (2018). Thus filter clips by a lightweight discriminability distillation network and only aggregate clips with high quality can be more efficient and practical to industry application.

Model	DDL (9 clips)		Random (9 clips)		Uniform (9 clips)		Dense (all clips)	
	Acc(%)	GFLOPs×clips	Acc(%)	GFLOPs×clips	Acc(%)	GFLOPs×clips	Acc(%)	GFLOPs×clips
3D-RS-50	86.38	37×9+1.7×60	82.83	37×9	83.14	37×9	83.92	37×60
R(2+1)D-RS-50	89.08	39×9+1.7×60	84.51	39×9	84.89	39×9	85.46	39×60
SlowFast-RS-50	**90.21**	16×9+1.7×60	85.92	16×9	86.14	16×9	87.72	16×60

Table 4: Video action recognition results on ActivityNet-1.2 dataset. We compare randomly, uniformly and filter by DDL to select 9 clips to aggregate. Accuracy is reported on the valiation set. And the compute consumption is estimated on a 2 minutes videos (above 60 clips), for that 2 minutes is the average video length of ActivityNet-1.2.

We test our DDL on untrimmed video action recognition with the ActivityNet-1.2 dataset. The ActivityNet-1.2 dataset contains 4,819 training videos and 2,383 validation videos for 100 action class. The duration of those videos vary from 2 seconds to 4 minutes and the average video length is 2 minutes. We carefully remove videos with multi-label. Frames are extracted from videos and resized the width to 240 pixels.

To combine and compare with DDL, we choose three clips-based 3D CNN models 3D-ResNet-50 Hara et al. (2018), SlowFast-50 Feichtenhofer et al. (2018) and R(2+1)D-50 Tran et al. (2018). To prevent over-fitting, we use the Kinetics-700 as pre-trained datasets. Video clips with adjacent 64 frames are randomly sampled from raw videos for training and testing. For the spatial domain, we randomly crop 112*112 pixels with a shorted side randomly sampled in [128, 160] pixels. 32 NVIDIA V100 GPUS with synchronized SGD training and global BN are used to train and we find that result for typical training in one 8 GPUs machine is the same. Similar to the 2D-DDL, we generate the pseudo quality label for each clip and use a channel reduction 3D resnet-18 network to regress the score after training the base model.

During inference, we choose top-K discriminability and aggregate them by average pooling. For comparison, we randomly or uniformly sample K clips in videos. K=9 is selected here to achieve accuracy and efficiency trade-off. Dense sampling is also used to compare, but it is not practical because of huge compute resource consumption.

As shown in table 4, DDL improves recognition performance for all baseline model. For the state-of-the-arts clips-based model SlowFast, combing with DDL achieves 4% accuracy gain. What's more, DDL can even outperform dense evaluation along with the whole video by a huge margin, which introduces dozens of times computation resource consumption. It is reasonable that DDL can select and aggregate clips with the highest quality, which avoids low-quality videos with visual blur or ambiguous content to pollute the video-level prediction.

Evaluation on Kinetics700.

we also evaluate our DDL on the popular trimmed action recognition benchmark Kinetics. Since around 2% videos of Kinetics-600 and 12% of videos of Kinetics-400 can't get access from youtube, we use the latest version Kinetics-700. Kinetics-700 contains 700 classes and 650k videos, and each action class has at least 600 video clips. Each clip is human annotated with a single action class and lasts around 10s.

The baseline models we select and the training procedure is similar to the last section besides the clip length is 32 frame. As shown in the table 5, DDL outperform random sample by 1.84 % and uniform sample by 2.18 %. For dense sample, DDL can achieve 0.86 performance gain with 6x speed up. Since Kinetics-700 is well trimmed, the improvement is not so large compared to untrimmed video dataset ActivityNet-1.2. However,as a post-training module, DDL can cooperate with any clip-based 3D action recognition model and achieve a more efficient and accurate inference, which is very suitable for real-world video understanding applications.

Cross dataset and model evaluation.

In the last two sections, we train specialized quality network for different dataset and different base model respectively, to demonstrate the generalization ability of our DDL, we do the cross dataset and cross model evolution experiment. Action recognition model R(2+1)D for ActivityNet-1.2 dataset is combined with DDL trained with Kinetics-700 and base model SlowFast, the results are shown

Model	DDL (5 clips)		Random (5 clips)		Uniform (5 clips)		Dense (30 clips)	
	Acc(%)	GFLOPs×clips	Acc(%)	GFLOPs×clips	Acc(%)	GFLOPs×clips	Acc(%)	GFLOPs×clips
3D-RS-50	71.01	37×5+1.7×30	68.26	37×5	67.43	37×5	68.83	37×30
R(2+1)D-RS-50	72.51	35×5+1.7×30	69.24	35×5	68.79	35×5	70.94	35×30
SlowFast-RS-50	**74.23**	16×5+1.7×30	72.39	16×5	72.05	16×5	73.37	16×30

Table 5: Video action recognition results on Kinetics-700 dataset. We compare randomly, uniformly and filter by DDL to select 5 clips to aggregate. Accuracy is reported on the validation set and is the average of top1 and top5 accuracy. Dense sample strategy sample 10 clips along the temporal axis and random crop 3 clips on the spatial axis.

DDL trained with	Accuracy(%)
SlowFast-RS-50/ActivityNet-1.2	87.91
SlowFast-RS-50/Kinetics-700	86.31
R(2+1)D/ActivityNet-1.2	89.08
R(2+1)D/Kinetics-700	86.92
Uniform	84.51
Random	84.80
Dense	85.46

Table 6: Results for cross dataset and cross model experiment. All results are reported with base action recognition R(2+1)D-ResNet-50 on ActivityNet-1.2 dataset. The training procedures for DDL vary from dataset and base model.

in table 6. It can be found that through trained with different dataset and base model causes a performance drop, DDL can still improve about 1% accuracy for dense sampling. These results show that the quality pattern learned by DDL is robust and well to generalize to other datasets. What's more, the generalization ability can avoid the training quality model for each action model and each dataset, making the DDL more efficient and pratical.

To better explore the pattern learned by DDL, we visualize some pictures in the YouTube Face datasets and their quality score generated by DDL.

5 Conclusion

In this paper, we have proposed a novel post-processing module called DDL for all group-based recognition tasks. We explicitly define the discriminability with observations on feature embedding, then a light-weight network is applied for quality distillation and feature aggregation. We identify the advantage of our proposed methods in the following aspects: (1) The entire discriminability distillation is performed without modifying the pretrained base network, which is highly flexible comparing with existing quality aware or attention ethods. (2) Our distillation network is extremely light-weighted with great computational cost salvage. (3) With our DDL and feature aggregation, we achieved state-of-the-art results on set-to-set face recognition and action recognition tasks.

References

Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 67–74. IEEE, 2018.

Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *Chinese Conference on Biometric Recognition*, pp. 428–438. Springer, 2018.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019a.

Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641*, 2019b.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. *arXiv preprint arXiv:1812.03982*, 2018.

Rohit Girdhar, Deva Ramanan, Abhinav Gupta, Josef Sivic, and Bryan Russell. Actionvlad: Learning spatio-temporal aggregation for action classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 971–980, 2017.

Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*, pp. 87–102. Springer, 2016.

Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 6546–6555, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *Computer Science*, 14(7):38–39, 2015.

iQIYI. iqiyi-vid-face, 2019. URL `http://challenge.ai.iqiyi.com/data-cluster`.

Yu Liu, Hongyang Li, and Xiaogang Wang. Rethinking feature discrimination and polymerization for large-scale recognition. *arXiv preprint arXiv:1710.00870*, 2017a.

Yu Liu, Junjie Yan, and Wanli Ouyang. Quality aware network for set to set recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5790–5799, 2017b.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pp. 568–576, 2014a.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014b.

Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In *Advances in neural information processing systems*, pp. 1988–1996, 2014.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 4489–4497, 2015.

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 6450–6459, 2018.

Fei Wang, Liren Chen, Cheng Li, Shiyao Huang, Yanjie Chen, Chen Qian, and Chen Change Loy. The devil of face recognition is in the noise. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 765–780, 2018.

Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pp. 499–515. Springer, 2016.

Jiaolong Yang, Peiran Ren, Dongqing Zhang, Dong Chen, Fang Wen, Hongdong Li, and Gang Hua. Neural aggregation network for video face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4362–4371, 2017.

Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4694–4702, 2015.

Xingcheng Zhang, Zhizhong Li, Chen Change Loy, and Dahua Lin. Polynet: A pursuit of structural diversity in very deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 718–726, 2017.

A Appendix

A.1 Qualitative analysis

We visualize images in frames set with their discrimination scores. Frames within a video with different discrimination scores are first shown. As we can see in pictures, our DLL learns meaningful discriminative feature patterns. Frames that meets large pose, visual blur, semantic ambiguous are assigned to a low score and those frame with well discriminative appearance are high weight. Aggregate frames by weight sum with such meaning discrimination patterns can form better set representation.

Discrimination scores for frames from the different sequence sets are shown next, which indicates the generalized ability among image sets.

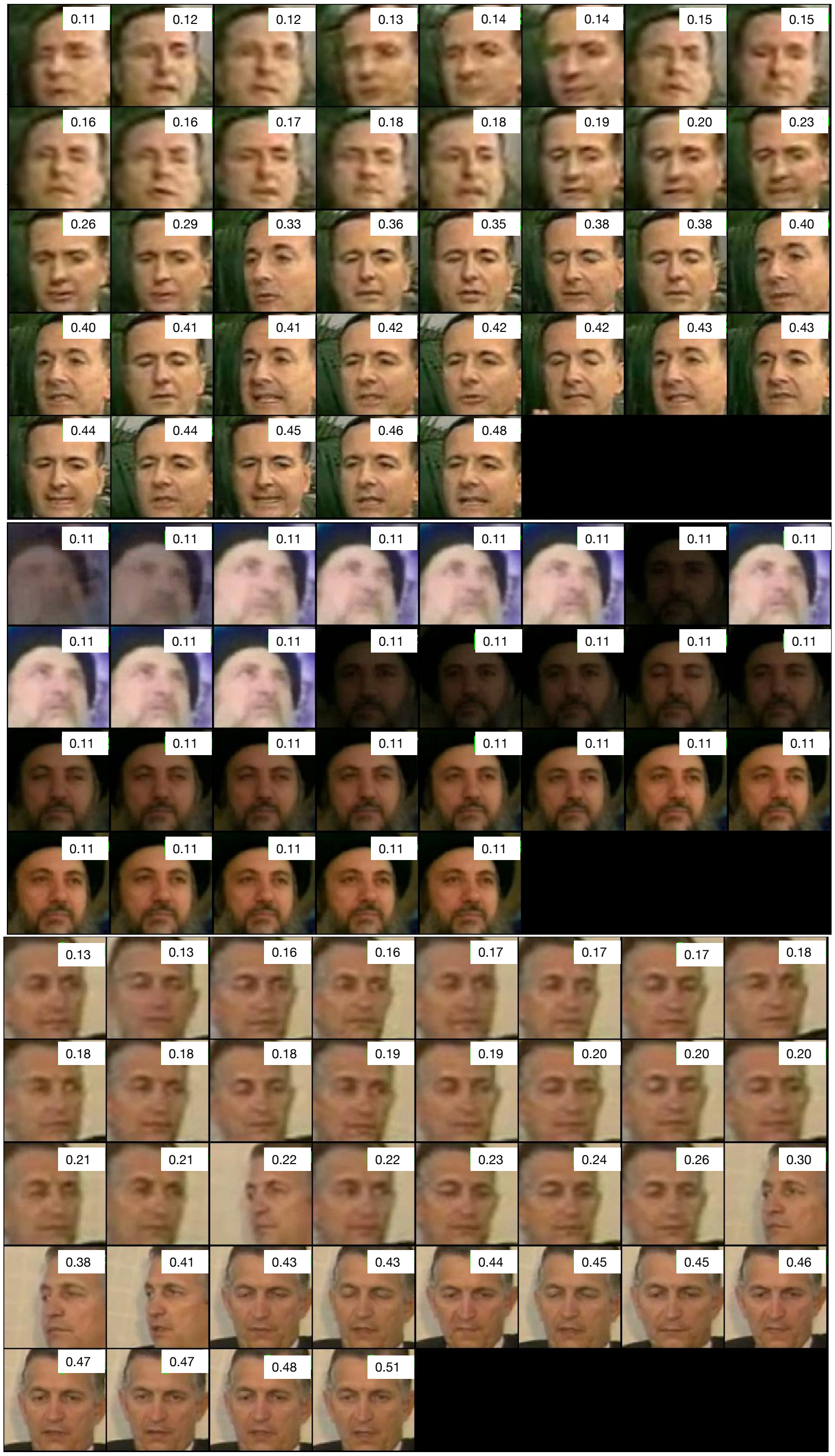

Figure 3: Visualization discrimiability of frames from one video.

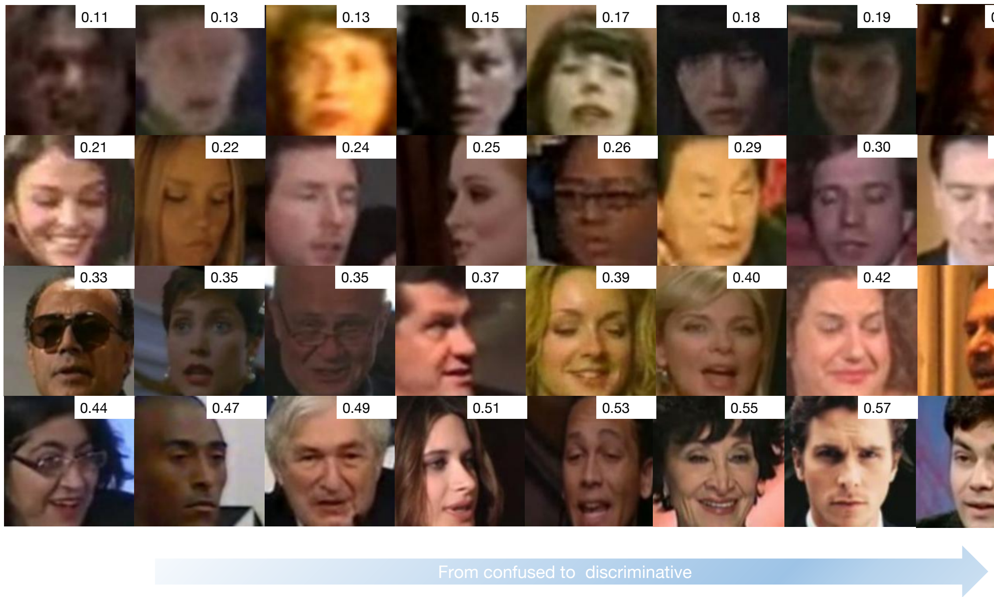

Figure 4: Visualization discrimiability of frames from different videos.